DOCUMENT RESUME

ED 175 236 FL 009 160

AUTHOR Musso, V.; Zangrando, R.
TITLE The Fiat Thesaurus: An Experiment by a Multinational Company.
INSTITUTION Commission des Communautes Europeennes (Luxembourg).
PUB DATE 77
NOTE 21p.; Paper presented at the European Congress on Information Systems and Networks (3rd, Luxembourg, May 3-6, 1977)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Artificial Languages; Business Communication; Computational Linguistics; Data Processing; Documentation; Indexes (Locaters); *Industry; *Information Retrieval; *International Trade Vocabulary; *Italian; Languages for Special Purposes; Reference Books; *Subject Index Terms; *Thesauri; Vocabulary

ABSTRACT

The vocabulary compiled and currently in use at "FIAT" comprises approximately 4000 words and is continually increasing. Some words are uniterms and others are polyterms selected during the scanning of documents of specific interest to the company. An attempt was made to retain a link with natural language by treating proper names and functors in a special way. Although the keywords contained in the vocabulary are mainly in Italian, other languages are also represented. The foreign terms adopted are those representing concepts which the Italian language does not yet cover. The vocabulary proved to be comprehensible to persons whose native language was not Italian, such as trainees of various nationalities who were employed by "FIAT." An experiment was carried out also with bibliographical documentary tapes available on the market. This experiment proved to be somewhat less successful mainly because instructions for use of the thesaurus were different for input and output. The guiding principles for future work are summed up as follows: deeper linguistic exploration of the problem, more assistance for users, and definition of the role of the thesaurus with batch or interactive processing. (Author/AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Vittorio Musso

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED175236

THE FIAT THESAURUS:

AN EXPERIMENT BY A MULTINATIONAL COMPANY

V.Musso and R.Zangrando

FIAT - Sèrvizi Consortili - Documentazione

Abstract

The vocabulary compiled and currently in use at FIAT comprises approximately 4000 words and is growing all the time. Some words are uniterms and others polyterms selected during the scanning of documents of specific interest to the Company. An attempt was made to retain a link with natural language by treating proper names and functors in a special way.

Comparison with similar experiments (OECD, ECJ and the like) has confirmed that the solutions adopted during compilation of the vocabularly satisfy wide-ranging requirements.

We have deliberately avoided applying preconceived ideas, preferring a more pragmatic approach free from obstructions hampering cooperation with the users. In fact, the reorganization of the Company has led to this unit taking on the role of central coordinator so that what was originally to be a specialized documentation centre now has a new primary task.

The keywords contained in the vocabulary are mainly in Italian, but other languages are also represented. The foreign terms adopted which have not been translated are those representing concepts which the Italian language does not yet cover (and in some cases these have been paraphrased).

The vocabulary stood up well when used by persons whose mother tongue was not Italian, such as trainees of various nationalities who were employed by us.

The performance of the bibliographical documentary tapes available on the market was, however, less satisfactory. The aspects requiring most attention were the instructions for use of the thesaurus which were different for input and output.

While some cooperation proved quite satisfactory, other attempts proved less useful as the user had wrongly assessed his real requirements and capabilities.

The guiding principles for our future work can be summed up as follows: deeper linguistic exploration of the problem, more assistance for users, definition of the role of the thesaurus with batch or interactive processing.

FL009140

1. A SHORT HISTORY OF THE COMPILATION OF THE FIAT THESAURUS

Fiat's experiment in this field started about 1971 and was prompted by the existence at that time of a considerable quantity of documentation ready for automatic processing. Original documents had in fact already been analyzed and processed, which involved the defining of their characteristic features and their conversion into standardized words which could be stored.

The choice of the characteristic features of a document also made us decide to use descriptors capable of describing the content of the document with a view to automatic retrieval, possibly also batch retrieval, and to achieve this not only on the basis of a particular arrangement or classification but also on the basis of a word which gave some indication of content by virtue of its semantic value.

Once production of these derived documents had reached a certain degree of consistency, we began to have doubts about the meaning attributed to the various words used as descriptors. It thus became essential to standardize and revise the descriptors which had been initially chosen more or less by instinct. In this context it should be pointed out that almost all our work is essentially an attempt to solve a problem, and by that we mean activity aimed at finding a solution in time with the aid of the knowledge gained from day-to-day experience.

Since the organization of the company's documentation centre where the work was being carried out obliged us to adopt a standardized language, not obtainable outside the firm, when we contacted the users, the documentalists were also obliged to develop a procedure for standardizing the words.

The various socio-economic thesauri available on the market (e.g. OECD, ILO etc), which were devised and organized to meet requirements which were not quite the same as those within our Company, convinced us that any documentation system must be customized to meet the requirements of the sector in

in which it operates.

Searching for documents can never be an objective activity since it is always influenced by the searcher's cultural background and the language used by him. It is thus essentially a subjective activity and as such assumes that the standardization of language, and consequently the wording of the thesaurus, also take account of environmental factors. Thus we cannot assume a priori that automatic translation of words into various languages is achievable, whereas an interpretation of ideas and concepts is. In our case the user was FIAT and thus we had to find a means which, in addition to playing a didactic role, could stimulate those involved to identify their real requirements.

It was for this reason that a thesaurus was compiled - a result of the need to put the words used as keys into logical order rather than a mere alphabetical list, putting the onus on the searcher to arrange the sequence of words each time.

Another reason for our decision in favour of independent processing was that external thesauri were more comprehensive but less relevant on certain problems since the degree of detail with which certain subjects were handled did not meet current requirements, which does not however affect the validity of the attempt to produce a generalized language. Day-to-day experience of work in a multinational company has led us to devote more attention to the development of languages connected with certain company divisions, such as personnel, production, patents, management.

We adopted the following approach: All the keywords used were extracted from the available derived documents by computers, and an alphabetical list was drawn up which was given to the documentalists who deleted anything which was no longer current or was ambiguous and reorganized the words in a logical way in accordance with the model shown in Annex 1, adding appropriate words where necessary, and polishing up, updating and completing the list. It should be noted that the procedure was based on existing derived documents and that this approach, as described by our Dr. Ghiglia in

Luxembourg, is similar to the one proposed by the ISO in its Standard 7288 on monolingual thesauri.

The work of extracting the descriptors by computer took about 4 days of programme preparation and the alphabetical list was a relatively short section of the programme, the input already being available on data carriers. Overall the documentalists spent a total of about 2 man-years within the space of 4 months on this project.

Once the various terms had been arranged in accordance with the model mentioned, an appropriate code number was allocated to each term (more about that later) and the programmes were produced with the derived lists as output which provided the dual-purpose list of the words used: alphabetical with code reference, and grouped by code, as is shown in Annexes 2 and 3.

As can be seen from the model, our main aim was to establish a system capable of handling a number of different subjects which were likely to be subsequently updated. Updating here means both the updating of the keywords within a given sector and the possible development of new subjects on the basis of the user's requirements.

We did not set out to compile a thesaurus which was valid in absolute terms for Fiat as a whole, but rather to define approaches and standards with which to create a thesaurus comprising sections which although diversified were coherent. The alternative would have been either an enormous and relatively inflexible thesaurus, or the neglecting of some subjects. This approach, which was designed to set up a thesaurus comprising sub-thesauri based on a common principle, in a sense anticipated the events which led Fiat to reorganize the company on the basis of a whole series of groups and independent units, responsible to a central controlling body operating with very simple parameters in the interests of uniformity.

For the thesaurus, we analysed the texts from the economic studies department and selected the main concepts. Then we translated these into keywords, thus creating an artificial language very close to natural language. As time went by we

began to realize that not all of the socio-economic fields were being adequately dealt with, and we therefore contacted internal and external departments of Fiat so that with their help we could fill the gaps.

However, not all those involved in the project were sufficiently interested in or prepared for the task ahead. They had a very good knowledge of their own sector from a strictly technical point of view but lacked linguistic experience, which meant that the approach adopted with regard to the thesaurus was not a standard one. Whilst not wishing to go into the technicalities of Italian or foreign words - the basic corpus of this thesaurus obviously being in Italian - we should like to give an account of some of our experience with the operation of this thesaurus.

2. NOTES ON THE USE OF THE THESAURUS

It should first of all be mentioned that the thesaurus was not used exclusively by Italian speakers, but also by foreigners. In our opinion the thesaurus fared quite well in its job of acting as a tool for the individual user with the aid of which he can describe the subject contained in the document using terms which have a definite meaning understood by all.

Some foreign users who had a fairly good knowledge of Italian had no trouble in grasping this somewhat artificial language because the underlying concept is so straightforward. Other foreign users who did not know so much Italian had more trouble. We then decided to accept reviews of documents in their mother tongue from which we extracted the main concepts and translated them, using the terms in our thesaurus.

This was an extremely effective way of introducing the foreigners to our thesaurus language. The thesaurus was also studied by a firm of foreign consultants who wanted to evaluate its structure and consistency. Our consultants pointed out that the thesaurus was designed more for batch processing than on line processing. The general structure

was defined as not only effective but also transferable to other languages with the use of the numerical code allocated to each keyword.

Unfortunately we have had no opportunity to up-date the thesaurus continuously as planned, at least to the extent we had intended. This would have involved setting up a central unit responsible for managing the thesaurus, which would record the requirements of peripheral units and be competent and equipped to update various items in the thesaurus. The organization of the company at that time did not allow for such a project to be started. Now, with reorganization of the company under way, we have an opportunity of a revising of the thesaurus along the lines envisaged, so as to cover also the semantic links between the various words.

Meanwhile we have carried out partial updating in conjunction with another firm, Siteco, which decided to delete a number of terms for its sector and modify others after a trial period during which it was indexing documents and attempting to retrieve them using the same descriptors. The thesaurus has also been used by persons who did not help to create it or review any documents, and in cases such as these the thesaurus has worked well, helping to formulate queries for the searches.

This point has been made because it might appear that the thesaurus was only of use to persons who were acquainted with its inner arrangement. In fact, even with users who were not particularly trained in methods of documentary searching, this organized list of words helped to put more exact queries.

The following notes will show how we intend to update the thesaurus and what we have done so far in conjunction with Siteco.

a) Siteco selected words from our thesaurus which it considered relevant to its own sector.
b) Together we chose words relevant to the Siteco sector from those contained in a list of words to be incorporated into our thesaurus.

c) Siteco suggested the addition of new words which were relevant and would not be otherwise retrieved.
d) The words from the three different sources were used to process the new sub-thesaurus for Siteco.

Once the documentalists and searchers had started to use this thesaurus a subsequent check showed that many of the words selected from other sources deliberately (see item c) had been used neither during indexing nor in the searches.

3.0 TECHNICAL NOTES

From the appended general table it can be seen that the thesaurus is arranged in three levels, even if for operational purposes only two are ever used. Level one is given by a code in two sections: the first part contains the first character denoting the subject and the second two characters denoting the chapters; the second section denotes the individual keywords.

The codes are in fact arranged on two levels; the subdivision into three components is purely an operational device and does not correspond to the actual function either in indexing or in retrieval. While the thesaurus was being compiled, however, it seemed a good idea to group certain subjects on the basis of similarity of meaning so that it would be easier to consult the thesaurus (see Annex). As far as the actual use of the thesaurus is concerned, however, the selection made reflects an attempt to obtain as much detail as possible in the indexing, whilst remaining free to group concepts or groups of concepts in various ways. Only keywords are used for indexing, while during the search, chapters can also be used. The latter procedure is equivalent to searching by using the logical sum of all the descriptors contained in the chapter.

3.1 The identification of a word with the aid of a code which also defines its dependence on wider concepts (chapter/descriptor) has another particularly important advantage for the

subject we are dealing with.

One of the major problems affecting thesauri is the establishment of a link between words in different languages. We resolved to link words in different languages with the aid of the code by giving them the same meaning. For example, if we consider the term Italia, in English 'Italy', in French 'Italie', by giving the same code to these three words a relationship of synonymity is established between the words, and during the search it is possible to change from any one of the terms to the other two. A possible immediate objection is that when horizontal - as well as vertical - links are established between the contents of the various words, it could be very difficult to keep track of changes in the meaning of a word if the code links are managed automatically by computer.

The first stage in the production of the thesaurus consisted in distinguishing the various types of descriptors, e.g. the functors, proper names, dates and descriptors properly speaking. We understand a functor as being a term which in itself does not have an unambiguous search meaning, but is used rather to clarify and guide the use of the actual keywords during the search. Proper names also were entered on the list of keywords, but no links were established between them or with other words in other chapters, consistency of presentation is the only reason why they were included in a special, differently structured chapter of the thesaurus.

Particular procedures, which are still being studied, were necessary for incorporating dates in the thesaurus. At the moment they are grouped in a single chapter, also organized in a non-standard way.

The descriptor is not always a uniterm, and in fact the polyterm descriptors represent perhaps the most original part of the FIAT thesaurus. Functors and dates can be shared with other thesauri, while the polyterms are concepts which purely reflect company sectors and as such represent company activities.

During the indexing of the documents, the expression of a

concept by means of a number of uniterms or of a definitely smaller number of polyterms might suggest the adoption of two different approaches, whereas in reality it is one single approach. Let us imagine that we are confronted with a series of concepts which all have something in common, and let us compare them for example to the sum of the powers of a single number. As in the case of algebraic procedures, we show the common element of these numbers, leaving in brackets a sum of prime numbers which cannot be further divided. The common element does not necessarily have a meaning of its own.

The polyterm descriptors, in other words the concepts we regard as being reducible to a sum of uniterms linked up in various ways, but which have a unambiguous meaning, are, so to speak, the prime numbers, i.e. the numbers which are not further reducible and thus had to be adopted as they were formed.

The OECD thesaurus has also adopted polyterm and uniterm expressions, basing the choice on the use or frequency of use of the descriptors formed. Like our thesaurus, it includes descriptors grouped according to aspects and large subject fields. Here too, the numerical code of the aspects provides a cross-reference between the alphabetical part and the part based on groups. Where interest profiles are needed, a particular number given by the computer to each word could be used instead. The technical thesaurus of the ECJ, which we studied some time ago to discover whether it was compatible with our system, also has a similar structure. We tried to find a link between the numerical code of our keywords and that of the universal decimal classification. We realized, however, that this was not wholly feasible since the numerical code used by the UDC can in fact indicate a range of concepts, and this is something which is not allowed for by the structure of our thesaurus. Partial agreement was possible, however, in that the first part of the UDC code could be adopted, in other words, the part which refers not to a single keyword as such but to the chapter containing it.

3.2 Measuring instruments

Another attempt at analysis was made on the individual words, used either as uniterms or as pluriterm descriptors. Often the functors were involved, in other words the words used to make the meaning of special expressions unambiguous.

We examined how these words were repeated in the vocabularly itself. For example: the term 'analisi' appears in a number of meanings (analisi statistica, analisi di mercato, analisi monetaria). The programme showed how many times this term was used in the various descriptors and also its distribution through the chapters. This was a first attempt to obtain a numerical indication of the words in our thesaurus, and it showed that a relatively limited number of words can be used to produce quite a large number of polyterm descriptors.

3.3 Links between inquiries and the thesaurus

It is the job of a thesaurus to supply the interface between the free or natural language of the question put by the searcher and the formalized language typical of some documentary stores. Our SDI system, which was carried out on the basis of a pilot project, is a typical example of this function as performed by the thesaurus. In the course of a fact-finding meeting with the system user, our expert documentalist tried to formalize in free speech and as faithfully as possible all the documentary interests of the user. Where the interests were many and varied, they were grouped into individual sectors. Then without further involvement of the user, the interest profile was 'translated' into thesaurus words, i.e. given descriptors and a chapter. The appropriate numerical code was used for each descriptor and for each chapter. These codes are arranged as a Boolean algebraic expression which can be easily read by the automatic system.

Each time the store was updated, the system retrieved the documents using the expression given. The user was invited

to assess the replies received. On the basis of these replies, we could either modify the expressions already formulated or leave them unchanged.

We intended to take the automatic system a stage further by arranging for automatic codification of the descriptors themselves.

4. EXTERNAL TAPES

Another experiment we carried out was with bibliographical tapes. First of all it should be noted that almost all bibliographical tapes are in English and this could pose a problem for persons whose mother tongue is not English. In our case the corpus of the thesaurus is in Italian, and we were thus confronted not only with simple problems of conversion, of reading tapes from different machines, but also with the problem of the transcription of the individual documents to fit our format so that the use of the document was possible in each of the necessary procedures. Another type of difficulty encountered was the translation from a foreign language into Italian, or rather into a formalized language such as that represented by the words in our thesaurus. The arrangement of the document itself was another obstacle. Writers in different fields have their own different ways of tackling technical subjects, and this means that the documentation produced internally even if sometimes too personalized and hence of doubtful use, always reflects the basic approach of the producer. It could also be argued that this excessive personalization of the document could make it less objective and hence more difficult to use.

While tackling the problem of tapes produced abroad, we found ourselves confronted both by documents quite different from what we were used to -not only from the point of view of language, but mainly from the point of view of the form given to the concepts - and by a structure geared to requirements which were new to us.

Since we obviously could not convert the document (i.e.

not merely translate it) to make it compatible with our system, we have tried to modify our mental approach so that we could better understand and interpret the material offered. As an example of our thinking on this subject, we could say that it is better to study German in order to read Goethe in the original than to use an Italian translation, no matter how excellent.

With this aim in view, a multilingual thesaurus is doubtless the best interface allowing the searcher access to a store processed in an environment different from our own from the point of view of locality, and especially, of culture.

I S R A D / P 2

CAMPIONE DI OUTPUT

vocabolario socioeconomico - lista per codice

+	I	+	0205018	+	CONTROLLO DELLA GESTIONE	+	+
+	E	+	0205019	+	NON NOTIFICATION FACTORING	+	+
+	E	+	0205020	+	CONVENTIONAL FACTORING	+	+
+	E	+	0205021	+	FACTORING	+	+
+	I	+	0205022	+	FACTORING INTERNAZIONALE	+	+
+	E	+	0205023	+	MATURITY FACTORING	+	+
+	I	+	0205024	+	ASSISTENZA GESTIONALE	+	+
+	I	+	0205025	+	AUTOGESTIONE	+	+
+	I	+	0205026	+	PARTECIPAZIONE OPERAIA	+	+
+		+		+		+	+
+		+		+		+	+
+	I	+	0206001	+	CAPITOLO ECONOMIA IMPRESA	+	+
+		+		+		+	+
+		+		+		+	+
+		+		+		+	+
+	I	+	0206002	+	PRODUTTIVITA' GLOBALE	+	+
+	I	+	0206003	+	SVILUPPO AZIENDALE	+	+
+	I	+	0206004	+	INNOVAZIONE AZIENDALE	+	+

I S R A D / P 2 CAMPIONE DI OUTPUT

vocabolario socioeconomico - lista alfabetica

0524010	MODELLO DI REGRESSIONE	-	LINEARE	
0522010	PROGRAMMA	-	LINEARE	
0405016	PROGRAMMAZIONE NON	-	LINEARE	
0524011	MODELLO	-	LINEARE	GENERALE
0103014			LINEA	DI PRODUZIONE
0732003			LINEA	ITALIANA
0002125	UNITED STATES	-	LINES	
0223013	MARKETING AND PRODUCT	-	LINE	EVALUATION
0727007	ON	-	LINE	SIMULATION
0727010			LINGUAGGIO	SIMULATO
0750003			LINGUA	PARLATA
0750004			LINGUA	SCRITTA
0750002			LINGUISTICA	
0750005	EDUCAZIONE		LINGUISTICA	
0750001	CAPITOLO PROBLEMI		LINGUISTICI	
0020001			LIN - PIAO	

FIAT THESAURUS GENERAL STRUCTURE

TOPICS	CODE		CHAPTERS CODE		CODE	
T O P I C S	0	Generalities service words	00 001	Italian Industries	10	People names
			01		11	
			02 001	Foreign Industries	12	
			03		13	
			04 001	International Bodies	14	
			05		15	
			06 001	Italian Bodies	16	
			07		17	
			08 001	Foreign Bodies	18	
			09		19	
	1	Production connected activities	00 001	Industry	10	
			01 001	Industrial properties	11	Natural resources
			02 001	Production activities	12	
			03 001	Plants equipment	13	Mining Petroleum
			04		14	
			05 001	Productivity	15	
			06		16	
			07		17	
			08		18	
			09		19	

Language code	TOPIC CODE	CHAPTER CODE	CHAPTER / DESCRIPTOR CODE		
	1	01	001	Proprietà industriali	Chapter
	1	01	002	Brevetto comunitario	Descriptor
	1	01	003	Licensing	Codes in
	1	01	004	Proprietà industriali	The chapt.
	1	01	005	Segreto industriale	

FIAT THESAURUS — OBJECTIVES

- Guidelines to implement a thesaurus in the FIAT users community

- Specifications of the real needs in the FIAT users community rayloring the answer to them

- Consistency in the implementation of specialized sections

- To suit the general process of decentralization running in FIAT

Pragmatic approach

Starting point and existing metadocuments (about 15.000)

Language :

also foreign words	
service-keywords	set
proper nouns	set
dates	set
descriptors	set

The code (6 digits) is the pivot to logically locate the descriptors and to define all the liaisons among them

FIAT THESAURUS — D A T A

4030 DESCRIPTORS

MONOTERMS 1597

MULTITERMS 2433

AVERAGE LENGHT OF EACH WORD = 17 CHARACTERS

TIME

DOCUMENTALISTS 2 man-year

PROGRAMMERS 1 man-week

on a 4 month period to collect, analyze and define descriptors

PROGRAMMERS 1 man-month

on a 2 month to implement the procedure

FIAT THESAURUS	USE	
	INDEXING	BATCH RETRIEVAL
TOPICS		
CHAPTER		X
DESCRIPTORS	X	X

. Foreign stagiaires

. Bibliographic tapes

. Pilot project

. Sperry-univac find macrochannels

. Programs for data collections and some simple statistical analysis

. The procedure detects the "new" descriptors at each new input of metadocuments